Research - AI Model Routing Research

Alpasan, Lois-Kleinner

2026

Abstract

Research on multi-provider AI model routing.

AI Model Routing Research

Overview

Research on multi-provider AI model routing.

Key Papers

Title	Author	Year
"RouteLLM: Optimizing LLM Routing"	Stanford	2024
"LLM Routing with Cost Constraints"	MIT	2024
"Multi-Provider AI Gateway Design"	Microsoft	2024
"Model Selection for Cost-Effective Inference"	Google	2023

Routing Strategies

Cost-Based

```yaml
strategy: Select cheapest model meeting quality threshold
metrics:
  - Cost per 1K tokens
  - Quality score (human eval)
  - Latency SLA

implementation:
  - Maintain model cost database
  - Quality benchmarks per task type
  - Dynamic provider selection
```

Latency-Based

```yaml
strategy: Select model with best latency within budget
metrics:
  - P50/P99 latency
```

```
  - Time to first token
  - Concurrency limits

implementation:
  - Real-time latency monitoring
  - Predictive latency modeling
  - Provider health scoring
```

Quality-Based

```
strategy: Route complex tasks to capable models
metrics:
  - Task complexity scoring
  - Model capability matrix
  - User preference

implementation:
  - Task classification
  - Model capability database
  - Fallback chain
```

Next

- Distributed Systems Research

See Also

Related research, architecture, and whitepaper documentation.

- Research Overview
- Architecture Overview
- Whitepapers
- Performance Research